Spatio-Temporal Epidemic Forecasting with Graph-Based Transformer

Mahmoud Ezzat[1*], Youssef Mohamed Malek[3], Tamer AbdelKader[2,1], Nagwa Badr[1]

[1*]Department of Information Systems, Computer and Information Sciences, Ain Shams University, Abbasya, Cairo, 11566, Egypt.
[2]Faculty of Computer Science and Engineering, Galala University, Ain Sukhna, 43511, Egypt.
[3]Department of Media Engineering and Technology, Faculty of Engineering, German University in Cairo, New Cairo, 11835, Egypt.

*Corresponding author(s). E-mail(s): mahmoud.abdelmobdy@cis.asu.edu.eg;
Contributing authors: youssef.malek@student.guc.edu.eg; tamer.abdelkader@gu.edu.eg; tammabde@cis.asu.edu.eg; nagwabadr@cis.asu.edu.eg;

Abstract

Epidemic forecasting plays a vital role in modern public health systems. COVID-19 epidemic outbreak underscores the critical need to develop forecasting models that are both accurate and responsive. Recent research has demonstrated the potential of spatio-temporal deep learning (DL) models that integrate human mobility networks with Graph Neural Networks (GNNs) to forecast disease spread across regions. Existing models still face challenges in completely capturing complex, non-linear temporal dynamics and modeling long-range spatial dependencies. Integrating GNNs with Transformers has shown significant potential over recurrent DL models in epidemic forecasting. To bridge this gap, we introduce two novel spatio-temporal architectures that combine GNNs with Transformer-based temporal modeling. We propose a local attention model, which restricts self-attention to temporally adjacent windows of the same city/province node, and a global attention model, which leverages full sequence-wide attention across all cities/nodes to capture long-range temporal and spatial dependencies. We benchmark our approaches against a Graph Convolutional Recurrent Network (GCRN) baseline using two real-world datasets, a high-quality dataset in Spain and a Brazilian dataset. Our results demonstrate that both models outperform the baseline. In the Brazilian dataset, the data cleaning process significantly improved the GCRN's average Root Mean Square Error dropping from 63.71 to 27.43 and its Mean Directional Accuracy (MDA) more than doubled. The global attention model achieved the highest MDA in Barcelona (64.29%), demonstrating its strength in capturing directional shifts. In the local attention model, we found a significant reduction in Symmetric Mean Absolute Percentage Error (55.3%) and +21% MDA improvement over GCRN, confirming its robustness to capture both magnitudes and turning points. Moreover, our experiments highlight the superior capacity of Transformers to model complex spatio-temporal dependencies and enhance precision in public health decision making.

Keywords: COVID-19, graph convolutional network, transformer, attention mechanism, spatio-temporal

1 Introduction

The increasing frequency and severity of epidemic outbreaks underscore the critical need for accurate predictive models capable of forecasting disease spread. Such models enable governments to inform public health policy and optimize the allocation of resources for disease control and prevention. The emergence of COVID-19 has intensified this demand for advanced modeling approaches that can help mitigate the spread of the virus and predict future case counts. Timely and reliable forecasts are crucial, as early intervention measures can save countless lives. One of the central challenges in controlling pandemics is predicting their future trajectory, which highlights the necessity of time series analysis in epidemiology, where historical infection data are leveraged to model and forecast future outbreaks.

Traditional statistical models, such as the Susceptible-Infected-Recovered (SIR) and Susceptible-Exposed-Infected-Recovered (SEIR) frameworks, have long provided a foundation for understanding disease transmission dynamics and emphasized the value of mathematical formulations for outbreak forecasting. However, these models often struggle to capture the complex, non-linear, and dynamic nature of epidemic spread, especially when handling large volumes of temporally and spatially resolved mobility and epidemiological data. To address these limitations and meet the evolving demands of epidemic prediction, researchers have increasingly turned to advanced machine learning and DL methods.

Human mobility data plays a pivotal role in modeling disease spread, as it captures key dimensions of how individuals or groups move through space. Analyzing these patterns provides crucial insights into transmission pathways and informs predictive models of future outbreaks. Yet, modeling human mobility remains challenging due to its non-linear and often irregular nature, the presence of long-range dependencies in sequential movement, and the heterogeneity and sparsity of data collected from diverse sources. These challenges have motivated the development of multi-modal models that integrate multiple data types for a deeper understanding of movement dynamics. In this context, combining graph-based models with Transformer architectures has emerged as a promising strategy to tackle these complexities.

The adoption of Transformer models marks a notable shift in spatio-temporal modeling, offering significant improvements over traditional methods, particularly in the context of epidemic forecasting. Unlike conventional approaches that rely on rigid assumptions and simplistic interactions, Transformers use attention mechanisms to dynamically identify and weigh the most relevant input features, enabling the model to focus on the factors that most strongly influence disease spread. This capacity allows for a more nuanced representation of origin-destination (OD) flows and facilitates the integration of diverse data sources and complex interdependencies. More broadly, advanced DL techniques have been increasingly applied to OD flow modeling in various domains, demonstrating the evolution and versatility of modern analytical frameworks [1]. Notably, Transformer-based architectures have achieved state-of-the-art performance in forecasting [2] and have also been successfully applied to anomaly detection tasks [3]. These strengths make Transformer-based models highly suitable for addressing the intricacies of epidemic prediction.

GNNs further advance the modeling of epidemic dynamics by capturing structured relational information and complex spatial dependencies [4]. When integrated with sequential temporal models, such as Recurrent Neural Networks (RNNs) and Transformers, GNNs enable the joint learning of spatial and temporal dynamics by aggregating signals from neighboring nodes and evolving node and edge features over time [5]. Additionally, GNNs can seamlessly incorporate multi-modal data, including mobility patterns, demographic profiles, and intervention policies, providing a more comprehensive foundation for robust epidemic forecasting. Beyond epidemiology, GNNs have shown strong performance in general time series tasks such as forecasting, anomaly detection, classification, and imputation, underscoring their growing significance in spatio-temporal machine learning [6].

Recent spatio-temporal learning advances provide concrete design cues that motivate our approach and help situate its strengths. First, several works demonstrate that pairing graph-based spatial encoders with attention-enhanced temporal modules can better capture heterogeneous regional effects and long-range dependencies. For example, space-aware Transformer–recurrent hybrids (e.g., *SSGCRTN*) enrich graph-convolutional recurrent backbones with attention while injecting *space-specific* components to model regional heterogeneity [7]. Multi-view dynamic fusion architectures (e.g., *MSTDFGRN*) integrate complementary signals and update spatial relations over time within a GCRN-style framework, highlighting the benefits of *dynamic* graph formulation [8]. Causal and self-learned adjacency approaches (e.g., *PSTCGCN*) show that learning layer-wise graphs can capture multi-level spatial dependencies beyond fixed, mobility- or geography-derived edges [9]. Dual-scale interaction networks (e.g., *SDSINet*) emphasize that *multi-scale* spatio-temporal interactions matter for accuracy, especially when local and global trends coexist [10]. Finally, time-varying graph families, often discussed alongside time-invariant/time-dependent decompositions (*TIID*/*TVGCN*-style formulations), illustrate that disentangling stable structure from transient, time-specific relations can be beneficial when contact and travel patterns evolve [11].

Taken together, these lines of work converge on three themes that our design embraces: (i) **attention-driven temporal modeling** to handle long-range, irregular dynamics; (ii) **graph adaptivity** (or compatibility with adaptive maps) to reflect evolving inter-regional connectivity; and (iii) **multi-scale reasoning** to reconcile local fluctuations with broader importation-driven trends. Our proposed Transformer–GCN variants inherit these strengths by preserving graph-based spatial aggregation while replacing the recurrent temporal stack with attention, thereby enabling selective, context-dependent temporal integration without discarding the inductive bias of spatial message passing.

Moreover, our study is intentionally positioned as a *controlled comparison* that builds directly on a GCN+RNN baseline (e.g., GCRN-style models): we keep the spatial encoder family fixed (GCN) and vary only the temporal mechanism (RNN vs. Transformer). This isolates the contribution of attention-based temporal encoding under identical preprocessing, splits, and metrics, helping clarify when and why attention helps for epidemic forecasting with mobility-aware graphs. While complementary innovations such as fully learned or dual-scale graphs can be layered on top, our focus on the temporal swap provides a clear, reproducible test of a central hypothesis raised by recent literature.

The main contributions of this work are summarized as follows:

- We propose two novel Transformer-GCN architectures: the Local Transformer-GCN (LTGCN) and the Global Transformer-GCN (GTGCN).
- We conduct extensive experiments comparing our models to a strong GCRN baseline on two real-world datasets, demonstrating the superior performance of Transformer-based models in trend detection Mean Directional Accuracy (MDA) and magnitude calibration Symmetric Mean Absolute Percentage Error (SMAPE).
- We provide a comprehensive preprocessing pipeline, including normalization, to ensure data quality and enhance the generalization capability of our models.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on epidemic prediction. Section 3 details the proposed LTGCN and GTGCN models, including their architectures, problem formulation, preprocessing pipelines, and evaluation metrics. Section 4 describes the experimental setup and presents our results. Section 5 discusses key findings, highlights strengths and limitations, and outlines future research directions. Finally, Section 6 concludes the paper.

2 Related Work

The emergence of data analytics and computational techniques has significantly revealed our ability to predict epidemics, highlighting the need for new solutions to predict the evolution of outbreaks and the transmission of diseases. Traditionally, epidemic prediction models have relied on more basic statistical methods such as SIR and its extension SEIR. To implement better interventions and responses, modified SEIR models have been developed to model the COVID-19 epidemic, analyze its dynamics, and provide a framework for using mobility data [12, 13, 14]. These compartmental models have some limitations, such as estimating transmission parameters based on static assumptions and detailed statistics that are costly and resource intensive to collect. The Global Epidemic and Mobility Model (GLEaM) incorporates both compartmental dynamics and real-world human mobility networks to capture the spatial dimension of infectious disease spread [15]. However, it still depends on predefined parameters and static mobility assumptions, which may not reflect real-time behavioral or policy changes [16].

Deep learning techniques have emerged as powerful tools in epidemic forecasting, integrating both spatial and temporal dynamics to improve predictive accuracy. Many researches integrate DL techniques with compartmental models to reduce data dependency while estimating transmission parameters , thus improving the accuracy of epidemic forecasting [17, 18]. Another hybrid approach is the Epi-DNNs model [19], which uses a neural network to express transmission parameters that are considered the coefficients of the compartmental model. Nadler et al. [20] integrate the Susceptible-Infected-Recovered-Dead (SIRD) model with the Long Short-term Memory (LSTM) neural network to estimate time-varying parameters during outbreaks and leverage LSTM's capacity for learning complex temporal patterns, thus improving the accuracy of epidemic prediction.

To overcome the limitations of hybrid and traditional compartmental models, researchers turned to use DL sequence models such as RNNs and their variants, LSTM and Gated Recurrent Units (GRUs). Due to their ability to avoid vanishing gradients in long sequences and learn long-term dependencies, LSTM-based models have become the most popular among DL sequence models in the domain of epidemic forecasting [21]. Shastri et al. [22] conduct a comparative case study of COVID-19 cases in India and USA based on the variants of LSTM; Stacked LSTM, Bi-directional LSTM and Convolutional LSTM. According to their findings, Convolutional LSTM outperformed the other variants in terms of prediction accuracy. In [23], the authors demonstrated that Convolutional Neural Networks (CNN) and Multivariate CNN outperformed LSTM and GRU models when forecasting with very few features and less amount of historical data. Otherwise, another study [24] showed that LSTM outperformed CNN and Multilayer Perceptron (MLP) while forecasting COVID-19 cases in Egypt for a week and a month ahead. These models do not adequately incorporate spatial dependencies, which are essential components in the accurate modeling of infectious disease transmission. Furthermore, they may exhibit a deficiency in interpretability, which creates challenges in extracting valuable insights for policy-making or epidemiological research.

GNNs have gained significant traction in spatio-temporal modeling owing to their power in capturing complex and long-range spatial relationships and dependencies. The Graph Attention-based Spatial Temporal (GAST) model, for example, utilizes graph attention networks (GATs) to simulate epidemic dynamics, showcasing enhanced predictive performance for the spread of both influenza and COVID-19 [25]. Likewise, the Metapopulation-based Spatio-Temporal Attention Network (MPSTAN) integrates multi-patch epidemiological insights into a spatio-temporal framework, thereby increasing forecasting precision by dynamically characterizing inter-patch interactions [26]. The nodes of a graph are represented differently according to the modeling approach utilized in epidemic forecasting. In temporal modeling, each node represents a specific time step, capturing the temporal dynamics of the epidemic.

For instance, Li et al. [5] proposed an integrated model between GCN and Transformer, where a Transformer encodes temporal information, and GCN decodes it to capture spatial dependencies. While in spatial modeling, each node corresponds to a geographical region. Edges represent mobility connections between regions [27], and node features include time series data of infection rates [28]. To address the challenge of limited training data, they utilize a meta-learning approach to transfer knowledge between countries [27]. Duarte et al. [29] combined GCN-based encoders and recurrent temporal models to improve the accuracy of forecasting epidemic spread using a Brazilian intercity mobility network. Although these models have been applied successfully to epidemic time series forecasting, their reliance on recurrent architecture imposed limitations such as adversity in capturing long-range temporal dependencies, their limited flexibility in learning global attention across nodes and time steps, and their poor scalability due to the sequential computation.

On the other hand, Transformer architectures have demonstrated considerable efficiency in modeling long-term temporal dependencies via self-attention mechanisms, particularly when integrated with learnable positional encodings. Nevertheless, they generally fail to incorporate spatial structures and inter-node interactions unless they are explicitly adapted for such purposes. Therefore, in this paper, we address this gap by proposing two novel variants of spatio-temporal Transformer architectures designed to capture epidemic dynamics over mobility graphs; LTGCN, where each city applies a Transformer encoder to its own historical case sequence independently, and GTGCN, where the Transformer is applied over a flattened sequence of all node-time tokens, enabling full global attention across both space and time. The local variant offers strong scalability and faster training, while the global variant excels in capturing long-range spatio-temporal interactions when computationally feasible.

3 Methodology

To forecast COVID-19 cases across regions, we design a spatio-temporal DL pipeline grounded in graph-based representations of human mobility and standardized epidemic trajectories. This section details our formal problem definition, dataset construction, preprocessing pipeline, architectural components and Training & Evaluation steps.

3.1 Problem Definition

We address the task of short-term regional-level epidemic forecasting by modeling the spatio-temporal dynamics of COVID-19 using both historical case counts and human mobility data. Given a graph of mobility interactions and recent case trajectories, the objective is to predict the number of new cases in each region for the next day.

Let N be the number of nodes (regions), T the historical input length, F the number of features per node (e.g., daily cases), and $H = 1$ the forecast horizon. The input tensor is $\mathbf{X} \in \mathbb{R}^{B\times T\times N\times F}$, where B is the batch size. The target is $\hat{\mathbf{Y}} \in \mathbb{R}^{B\times N\times H}$.

The model also receives:

- A directed graph $G = (V, E)$ with edge weights w_{ij} encoding average weekly mobility between nodes i and j.
- Static node features $\mathbf{Z} \in \mathbb{R}^{N\times D}$, such as normalized population.

The learning task is to approximate a function

$$\mathcal{F} : (\mathbf{X}, G, \mathbf{Z}) \to \hat{\mathbf{Y}},$$

which outputs region-wise forecasts of new case counts across the forecast horizon.

3.2 Datasets

We evaluate our models on COVID-19 cases and human mobility datasets from Brazil and Spain. The Brazilian graph includes 5,385 cities and ∼21k edges (after backbone filtering), built from intercity vehicle flow data [30] and normalized using 2022 census population. The Spanish dataset includes province-level mobility from Ministry of Transport and Sustainable Mobility (MITMA) [31] and COVID-19 cases from Datadista [32], scaled by provincial population. In both cases, we construct directed and weighted graphs representing weekly flows.

3.3 Data Preprocessing

Accurate spatio-temporal forecasting depends on well-structured temporal signals and informative spatial graph representations. To ensure data quality and consistency, we apply a unified preprocessing pipeline to both the Brazil and Spain datasets, which include COVID-19 case counts and inter-regional mobility flows. This pipeline encompasses data cleaning, normalization, and backbone graph filtering, following the methodology established in our baseline study [29].

3.3.1 Brazil Dataset

The Brazilian dataset comprises daily COVID-19 reports from over 5,000 municipalities (2020–2023). To ensure quality, we filtered cities with over 40 days of inactivity, extreme negative values (e.g., < -200), or statistical outliers exceeding $\mu + 10\sigma$ in 30-day windows. This yielded a cleaner subset of 1,305 cities for model training.

Normalization.

We merged multi-year case files and clipped negatives to zero. Two normalization strategies were applied:

- **Per-capita scaling:** Cases per 100,000 people.
- **Z-score normalization:** Applied per city to standardize temporal dynamics.

Mobility Graph.

Using the intercity flow data [30], we constructed a weighted mobility graph and applied the disparity backbone algorithm [33] to retain statistically significant edges, following the approach adopted in the GCRN baseline study [29]. Edges with $p_{ij} < 0.01$ (from the disparity backbone [33]) or among the top-5 strongest connections per node were retained to maintain robust local connectivity. The graph was encoded using PyTorch Geometric [34] with node features including population and centrality.

3.3.2 Spain Dataset

For Spain (52 provinces), we applied similar steps. Negative case values were clipped and standardized via Z-score normalization. The population was scaled with min-max normalization.

Mobility Graph.

The weekly travel matrices from MITMA [31] were cleaned, scaled, and filtered with the same p_{ij}-based backbone strategy. The final sparse graph encoded normalized travel volumes between provinces.

These pipelines ensured reliable, scale-invariant inputs and interpretable graph structure for downstream learning.

3.4 Model Architectures

3.4.1 GCRN Baseline (Recurrent GCN)

GCRN serves as our baseline model and is adapted from the architecture proposed by Duarte et al. [29]. It captures both spatial and temporal dependencies by integrating graph convolutional layers within a GRU framework, as originally introduced by Seo et al. [35] and further developed for spatio-temporal forecasting in works like Diffusion Convolutional Recurrent Neural Network (DCRNN) [36].

At each time step t, the model updates node-level hidden states $\mathbf{H}_t \in \mathbb{R}^{B\times N\times H}$ based on the current input $\mathbf{X}_t \in \mathbb{R}^{B\times N\times F}$, the previous hidden state $\mathbf{H}_{t-1}$, and the mobility graph G. This update is defined as:

$$\mathbf{H}_t = \text{GRUGCN}(\mathbf{X}_t, \mathbf{H}_{t-1}, G)$$

The GRUGCN operator uses gated mechanisms—reset, update, and candidate gates—each parameterized via graph convolutions (GCNConv). This approach has been widely applied in spatio-temporal tasks such as traffic flow [36, 37] and infectious disease forecasting [29, 38].

After processing the full sequence of length T, the final hidden state $\mathbf{H}_T$ is passed through a feedforward layer with ReLU activation to produce predictions $\hat{\mathbf{Y}} \in \mathbb{R}^{B\times N\times H}$, where H is the forecast horizon (typically 1 day).

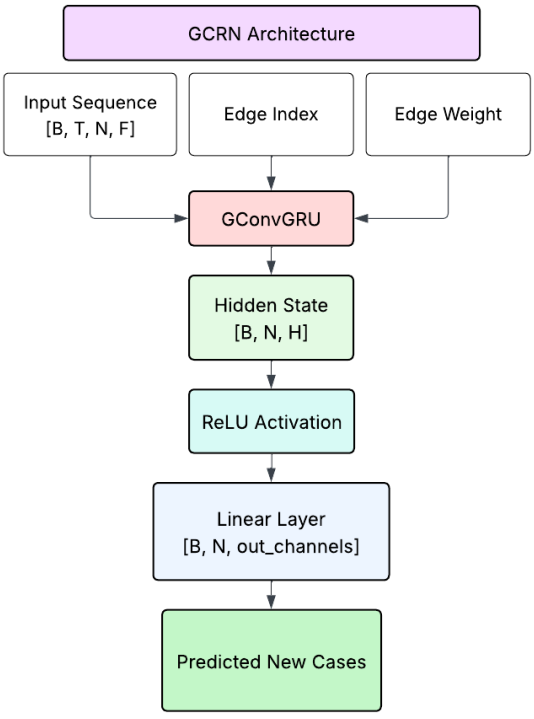

Fig. 1 Architecture of the GCRN baseline. Each input frame passes through a GCN-parameterized GRU cell, with final hidden states used to generate per-node forecasts.

We implement this architecture using a custom `GConvGRUCell` to support efficient batched training on sequences shaped $\mathbf{X} \in \mathbb{R}^{B\times T\times N\times F}$. An overview of the architecture is shown in Figure 1.

Although effective, this model is inherently sequential and thus limited in modeling long-range temporal dependencies or capturing complex cross-regional interactions beyond localized graph propagation.

Table 1 Architecture and hyperparameters of the GCRN model used in our experiments.

Component	**Configuration**
Input features	$in\_channels = 1$ (time-series feature: newCases)
Hidden state	$hidden\_channels = 512$ (per-node GRU state)
Output dimension	$out\_channels = 1$ (1-day ahead forecast per node)
Optimizer	Adam ($lr = 10^{-3}$)
Scheduler	StepLR ($step\_size = 5$, $\gamma = 0.5$)
	Model layers
Gated graph cell	GConvGRUCell with three GCNConv gates per step: z, r, and $\tilde{h}$
Per-step graph ops	For each $t \in \{1, \ldots, T\}$: concat $[x_t, h_{t-1}]$; apply GCNConv-based gates; update h_t
Readout	Linear($hidden\_channels \to out\_channels$) after the final h_T
Nonlinearities	σ for z and r gates; tanh for candidate; ReLU before final linear (optional)
Graph input	Shared $edge\_index$, optional $edge\_weight$ for all steps

As shown in Table 1, this configuration specifies the exact hyperparameters and architectural details we adopt for training the GCRN baseline in our experiments. These settings ensure consistency across runs and align with prior work in spatio-temporal graph forecasting.

3.4.2 LTGCN (Nodewise Temporal Attention)

To address the limitations of recurrent models in capturing long-range dependencies, we introduce the LTGCN model that replaces recurrence with Transformer encoders applied independently to each node's historical case trajectory. This design leverages the Transformer's self-attention mechanism [39] to enhance temporal expressiveness, building on recent advances in Transformer-based models for time series forecasting [2, 40, 41, 42]. An overview of the architecture and fusion mechanism is shown in Figure 2.

Temporal Encoder.

For each node i, we define the input sequence $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,T}] \in \mathbb{R}^{T \times F}$. This sequence is augmented with learnable positional encodings $P \in \mathbb{R}^{T \times d}$ and passed through a Transformer encoder:

$$\hat{X}_i = \text{TransformerEncoder}(X_i + P), \quad \hat{X}_i \in \mathbb{R}^{T \times d} \tag{1}$$

This formulation captures intra-node temporal dependencies, inspired by nodewise Transformer applications in spatio-temporal forecasting such as Spatial-Temporal Transformer Network (STTNs) [43]. .

Spatial Encoder.

To capture global spatial context, a static GCN is applied to node-level features. The spatial representation for node i is:

$$h_i^S = \text{GCN}(X_{\text{spatial}}, A)_i \tag{2}$$

where A denotes the adjacency matrix of the human mobility graph.

Fusion Attention.

To integrate temporal and spatial representations, we propose a local attention mechanism. First, a temporal summary $h_i^T \in \mathbb{R}^d$ is computed by averaging over the

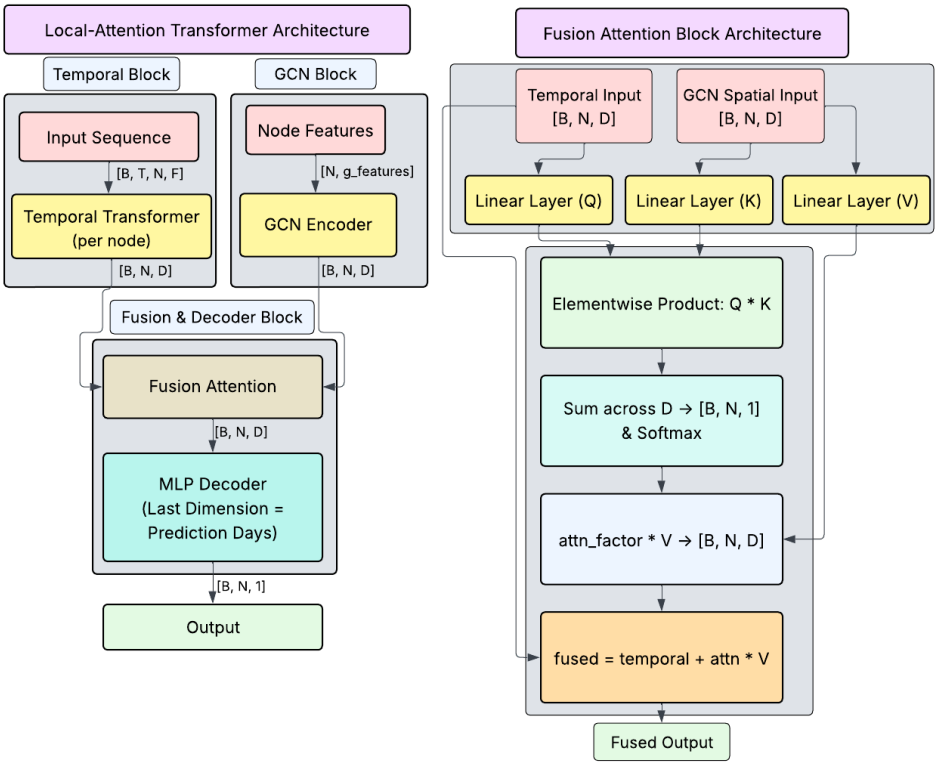

Fig. 2 Architecture of the LTGCN model. Each node's time series is encoded by a Transformer, while node embeddings are also computed globally via a GCN. A fusion attention mechanism integrates both temporal and spatial features per node.

Transformer output:

$$h_i^T = \frac{1}{T} \sum_{t=1}^{T} \hat{X}_{i,t} \tag{3}$$

The fused node representation h_i^{fused} is then computed using attention:

$$q_i = W_q h_i^T, \quad k_i = W_k h_i^S, \quad v_i = W_v h_i^S \tag{4}$$

$$\alpha_i = \text{softmax}(q_i \cdot k_i) \tag{5}$$

$$h_i^{\text{fused}} = h_i^T + \alpha_i v_i \tag{6}$$

where $W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ are learnable projection matrices. This attention mechanism selectively integrates spatial signals based on temporal dynamics, following joint encoder-fusion strategies commonly used in spatio-temporal forecasting models [44].

Model Characteristics.

By independently modeling each node's time series, the model provides fine-grained, long-range temporal forecasts. However, this design does not explicitly model inter-node temporal interactions, which can be critical for capturing how regional influences propagate during epidemic outbreaks, as demonstrated in recent spatio-temporal attention models [45].

As summarized in Table 2, this configuration defines the Local Spatiotemporal Transformer used in our experiments. The settings describe the precise hyperparameters and architectural choices employed for training, ensuring reproducibility and enabling a fair comparison against other baselines.

Table 2 Architecture and hyperparameters of the Local Spatiotemporal Transformer.

Component	Configuration
Input features	$in\_channels = 1$ (time-series feature: newCases)
Static node features	$graph\_feat\_dim = 1$ (e.g., population, centrality)
Hidden dimension	$trans\_hidden = 1024$ (used across GCN, Transformer, Fusion)
Output dimension	$out\_channels = 1$ (1-day ahead prediction per node)
Optimizer	Adam ($lr = 10^{-3}$)
Scheduler	StepLR ($step\_size = 10$, $\gamma = 0.75$)
Model layers	
Temporal encoder	Linear(1→1024) + Positional embedding + 1 Transformer layer ($nhead = 8$)
Spatial encoder	2-layer GCN: GCNConv(1→1024) → ReLU → GCNConv(1024→1024) → ReLU
Fusion	Attention-based fusion of temporal and spatial representations
Decoder	MLP: 1024→512→256→64→1 with ReLU activations

3.4.3 GTGCN (Flattened Global Attention)

To overcome the locality constraints of recurrent and node-wise models, we propose the GTGCN model. Unlike previous models that process each node independently or sequentially, this model treats each node-time pair as an individual token, allowing full spatio-temporal interaction across the entire input sequence. An overview is shown in Figure 3.

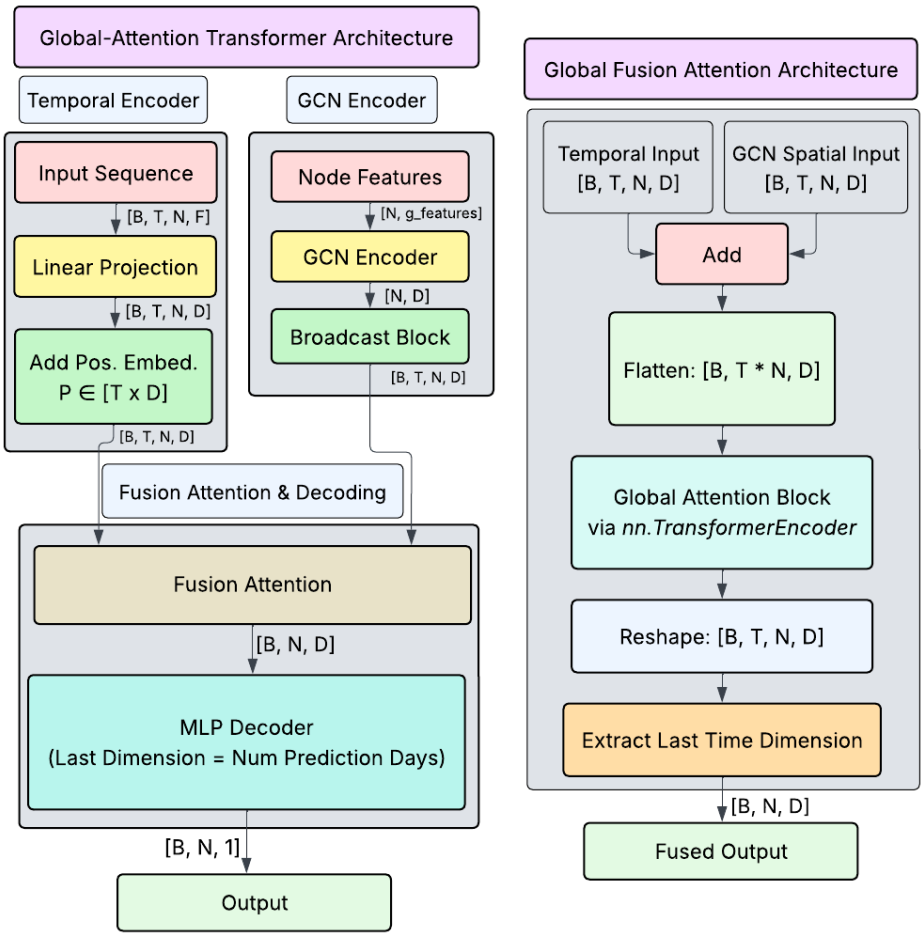

Fig. 3 Architecture of the GTGCN model. Node features are enriched using a GCN encoder and fused with temporal embeddings before being flattened into a global spatio-temporal token sequence. A causal Transformer encoder enables attention across both space and time.

Temporal-Spatial Tokenization.

The input $\mathbf{X} \in \mathbb{R}^{B\times T\times N\times F}$ is first linearly projected to a hidden dimension D, producing $\mathbf{X}_{\text{proj}} \in \mathbb{R}^{B\times T\times N\times D}$. Each node's static features $\mathbf{Z} \in \mathbb{R}^{N\times g}$ are processed via a GCN encoder:

$$\mathbf{S} = \text{GCN}(\mathbf{Z}, A) \in \mathbb{R}^{N\times D}, \quad \mathbf{S}_{\text{seq}} = \text{Broadcast}(\mathbf{S}) \in \mathbb{R}^{B\times T\times N\times D} \tag{7}$$

This follows prior work that encodes node-level attributes with GCNs for spatial reasoning [36, 46].

Positional Encoding and Fusion.

Learnable temporal embeddings $\mathbf{P} \in \mathbb{R}^{T\times N\times D}$ are added to encode time order. The fused input becomes:

$$\mathbf{X}_{\text{enc}} = \mathbf{X}_{\text{proj}} + \mathbf{S}_{\text{seq}} + \mathbf{P}$$

Similar temporal positional encoding strategies have been employed in time series Transformer models such as Informer [2] and Autoformer [41], though we adopt a learnable version here.

Global Transformer Encoding.

We flatten the 3D tensor to a sequence of shape $[B, T \cdot N, D]$, where each token corresponds to a specific node and time. A causal mask is applied to preserve autoregressive structure. A multi-layer Transformer encoder captures interactions across all tokens:

$$\mathbf{H} = \text{TransformerEncoder}(\text{Flatten}(\mathbf{X}_{\text{enc}}), \text{mask}) \tag{8}$$

This design enables nonlocal information exchange across the entire spatio-temporal space, inspired by global self-attention strategies in graph representation learning [47] and spatio-temporal forecasting [43].

Decoding.

After reshaping back to $[B, T, N, D]$, we select the final timestep $\mathbf{H}_T \in \mathbb{R}^{B\times N\times D}$ and pass it through MLP to predict case counts:

$$\hat{\mathbf{Y}} = \text{MLP}(\mathbf{H}_T)$$

Model Characteristics.

- **Global temporal-spatial attention:** Each node can attend to any other node's historical trajectory [44].
- **Learned positional embeddings:** Provides flexibility beyond fixed encodings [41].
- **GCN-enriched features:** Integrates static graph structure into temporal modeling [36, 48].

As presented in Table 3, this configuration outlines the Global Spatiotemporal Transformer used in our study. The table specifies the key hyperparameters and architectural components adopted for training, providing clarity on the experimental setup and ensuring comparability with other models.

This globally attentive Transformer design offers full parallelism and holistic epidemic reasoning, enabling the model to learn complex spatio-temporal interactions such as regional contagion spread patterns, lagged dependencies, and nonlocal outbreaks.

Table 3 Architecture and hyperparameters of the Global Spatiotemporal Transformer.

Component	Configuration
Input features	$input\_dim = 1$ (time-series feature: newCases)
Static node features	$gcn\_dim = 1$ (e.g., population, centrality)
Hidden dimension	$hidden\_dim = 1024$ (shared across modules)
Attention heads	$nhead = 32$
Transformer depth	$num\_layers = 1$ encoder layer
Optimizer	Adam ($lr = 10^{-3}$)
Forecast dimension	$forecast\_dim = 1$ (per-node scalar)
Dropout	Transformer = 0.2, Decoder MLP = 0.1
Model layers	
Input projection	Linear($1 \to 1024$)
Spatial encoder	Two-layer GCN: GCNConv($1 \to 1024$) $\to$ ReLU $\to$ GCNConv($1024 \to 1024$) $\to$ ReLU
Temporal encoding	Learnable time positional embeddings (length 1000)
Flattened tokens	Reshape to $[B, T \cdot N, 1024]$ with a causal attention mask
Transformer encoder	1 `TransformerEncoderLayer` ($d\_model = 1024$, $nhead = 32$, dropout 0.2)
Decoder head	MLP: $1024 \to 512 \to 256 \to 64 \to 1$ with ReLU, dropout 0.1
Output	$[B, N, 1]$ per batch

3.5 Training and Evaluation

We adopt a consistent training pipeline and evaluation strategy across both datasets to ensure fair model comparison and statistical robustness. This includes a sliding window forecasting setup, standardized loss functions, and multiple performance metrics.

3.5.1 Sliding Window Forecasting Strategy

Following prior work [29], we use a fixed-length sliding window approach to generate training samples. For each region (node), we extract:

- **Input window:** 14 days of historical case data.
- **Output window:** 1-day-ahead prediction.
- **Feature:** Z-score normalized daily new cases.

This approach produces a dense set of overlapping samples, increasing the availability of training data by more than 15× compared to non-overlapping strategies. Each sample is structured as $\mathbf{X} \in \mathbb{R}^{B \times T \times N \times 1}$, where B is batch size, $T = 14$, and N is the number of regions.

3.5.2 Optimization Setup

All models are trained using the Mean Squared Error (MSE) loss:

$$\mathcal{L} = \frac{1}{B \cdot N} \sum_{b=1}^{B} \sum_{n=1}^{N} (y_{bn} - \hat{y}_{bn})^2$$

This loss function is widely used in COVID-19 forecasting and spatio-temporal graph modeling due to its stability and interpretability.

We optimize the model parameters using the Adam optimizer [49] with an initial learning rate of 0.001. A step-based learning rate scheduler is applied to facilitate convergence:

- Brazil: learning rate decayed by 50% every 5 epochs.
- Spain: learning rate decayed by 30% every 5 epochs.

Weight decay (L2 regularization) is included for models requiring additional control over overfitting, as it is a standard strategy to improve generalization in Transformer-based and graph-based temporal models.

3.5.3 Evaluation Metrics

To assess both prediction accuracy and trend correctness in COVID-19 forecasts, we adopt three complementary metrics: Root Mean Square Error RMSE, SMAPE, and MDA. These metrics have been widely adopted in epidemic modeling and spatio-temporal forecasting [29, 38, 41].

Root Mean Squared Error (RMSE).

$$\text{RMSE} = \sqrt{\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\left(\hat{y}_{nt} - y_{nt}\right)^2}$$

RMSE quantifies the absolute magnitude error and is commonly adopted as a primary metric in COVID-19 forecasting benchmarks [29, 38]. It penalizes larger deviations more heavily, making it useful for assessing peak prediction reliability and extreme case dynamics.

Symmetric Mean Absolute Percentage Error (SMAPE).

$$\text{SMAPE} = \frac{100}{T}\sum_{t=1}^{T}\frac{|\hat{y}_t - y_t|}{(|\hat{y}_t| + |y_t| + \epsilon)}$$

SMAPE measures relative error while addressing the limitations of MAPE when actual values are close to zero. This is particularly relevant in epidemic forecasting, where case counts often drop to low levels in tail phases. Prior work has used MAPE to assess COVID-19 forecasting accuracy [5, 50, 51, 52], but we opt for SMAPE to ensure robustness in low-value regimes and avoid instability from small denominators.

Mean Directional Accuracy (MDA).

$$\text{MDA} = \frac{1}{T-1}\sum_{t=2}^{T}\mathbb{1}\left[(\hat{y}_t - \hat{y}_{t-1})(y_t - y_{t-1}) > 0\right]$$

MDA evaluates whether the model correctly predicts the direction (increase or decrease) of future cases, regardless of magnitude. This is crucial in epidemiological models to anticipate trend shifts, surges, or declines.

Training and Validation Loss.

In addition to the evaluation metrics above, we report the *training and validation losses per epoch* by providing their loss curves for all models. Since validation loss often fluctuates due to stochastic optimization and subgraph batching, we report the **Exponential Moving Average (EMA)** of the validation loss to highlight the underlying trend. The EMA reduces noise by weighting recent epochs more heavily:

$$\text{EMA}_t = \alpha\, L_t + (1-\alpha)\,\text{EMA}_{t-1},$$

where L_t is the raw validation loss at epoch t, and $\alpha \in (0, 1]$ is the smoothing factor (we use $\alpha = 0.2$ in our experiments). This smoothing highlights convergence behavior and mitigates spurious spikes while still reflecting meaningful trend changes.

3.5.4 Reconstruction and Visualization

Though models are trained on Z-score normalized inputs, we reverse the normalization for evaluation. Predictions $\hat{z}$ are reconstructed as:

$$\hat{y}_{\text{original}} = \hat{z} \cdot \sigma + \mu$$

where μ and σ are region-specific means and standard deviations. We visualize actual vs predicted trajectories for key regions (e.g., Brasília, Barcelona).

3.5.5 Evaluation Protocol

Each model is evaluated on a held-out test set composed of the final 20% segment of the time series. All metrics are computed at the regional and global levels. In addition to quantitative scores, we include detailed plots and error distributions to provide a holistic view of each model's behavior.

3.5.6 Model Input/Output Overview

We also provide a table below that demonstrates the various input and output dimension shapes per component in the various models and explanatory notes per each component.

4 Experiments

4.1 Experimental Setup

We compare recurrent and attention-based models for spatio-temporal forecasting across two settings. Our goal is to evaluate the impact of temporal encoders (GRU vs. Transformer) and spatial modeling (GCN), and to assess global vs. localized attention mechanisms under varying data quality.

Evaluation Metrics

In Section 3.5.3, we report RMSE, SMAPE, and MDA to measure absolute, relative, and directional performance, respectively.

4.2 Brazil Experiments

Table 5 shows that GCRN and the lighter LTGCN are trained on every Brazilian subset, while the GTGCN is restricted to the Top-40 cities because its $\mathcal{O}(N^2)$ attention cost becomes prohibitive on larger graphs. This layout lets us quantify the effect of data cleaning and graph size, yet still compare all three models on a tractable, high-priority subset.

Subset	**GCRN**	**LTGCN**	**GTGCN**
Unfiltered (5,300+ cities)	✓	✓	—
Cleaned (1,305 cities)	✓	✓	—
Top 40 cities	✓	✓	✓

Table 5 Coverage of models across Brazilian data subsets (✓ = model trained; — = not trained).

Table 4 Input and output dimensions of model components.
B = batch size, T = history length, N = number of nodes,
F = input features, H = forecast horizon, D = hidden dimension.

Component	**Input Shape**	**Output Shape**	**Notes**
Raw Input	$X \in \mathbb{R}^{B\times T\times N\times F}$	—	Daily case features per node
Target	—	$\hat{Y} \in \mathbb{R}^{B\times N\times 1}$	Forecast horizon (1 day)
GCRN (baseline)	$X_t \in \mathbb{R}^{B\times N\times F}$	$H_t \in \mathbb{R}^{B\times N\times H}$	GRU cell with GCN parameterization
LTGCN – Temporal Encoder	$X_i \in \mathbb{R}^{T\times F}$	$\hat{X}_i \in \mathbb{R}^{T\times D}$	Nodewise Transformer per region
LTGCN – Spatial Encoder (GCN)	$Z \in \mathbb{R}^{N\times D}$	$h_i^S \in \mathbb{R}^D$	Static node features (e.g., population)
LTGCN – Fusion Attention	$h_i^T \in \mathbb{R}^D, h_i^S \in \mathbb{R}^D$	$h_i^{fused} \in \mathbb{R}^D$	Combines temporal + spatial features
GTGCN – Projection	$X \in \mathbb{R}^{B\times T\times N\times F}$	$X_{proj} \in \mathbb{R}^{B\times T\times N\times D}$	Linear map to hidden dim
GTGCN – Spatial Encoder (GCN)	$Z \in \mathbb{R}^{N\times g}$	$S \in \mathbb{R}^{N\times D}$	Node embeddings broadcast across T
GTGCN – Flattened Tokens	$X_{enc} \in \mathbb{R}^{B\times T\times N\times D}$	$Seq \in \mathbb{R}^{B\times(T\cdot N)\times D}$	Each node-time pair as a token
GTGCN – Transformer Encoder	$Seq \in \mathbb{R}^{B\times(T\cdot N)\times D}$	$H \in \mathbb{R}^{B\times(T\cdot N)\times D}$	Global spatio-temporal self-attention
GTGCN – Decoder (MLP)	$H_T \in \mathbb{R}^{B\times N\times D}$	$\hat{Y} \in \mathbb{R}^{B\times N\times 1}$	Final forecast per node

4.3 Spain Experiments

With only 52 provinces and consistent data quality, we train all model types (GCRN, LTGCN, and GTGCN) on the full Spanish dataset. This serves as a robustness evaluation under cleaner epidemiological conditions and provides insight into attention scope vs. performance trade-offs.

4.4 Results on Brazilian Dataset

Quantitative Evaluation

Table 6 compares the forecasting performance of GCRN, LTGCN, and GTGCN across three subsets of the Brazilian dataset: the complete set of cities (5,300+), a cleaned high-quality subset (1,305 cities) and the top 40 most populous cities.

Subset	Metric	GCRN	LTGCN	GTGCN
Top 40 Cities	RMSE (Brasília)	524.88	**481.55**	571.37
	SMAPE (Brasília)	55.43%	**63.49%**	59.37%
	MDA (Brasília)	29.55%	**42.27%**	34.68%
	Avg. RMSE / 100k	26.55	26.22	**14.77**
	Avg. SMAPE	29.96%	30.35%	**26.91%**
	Avg. MDA	24.77%	**37.89%**	32.08%
Cleaned Cities	RMSE (Brasília)	527.42	**508.69**	—
	SMAPE (Brasília)	56.87%	**57.98%**	—
	MDA (Brasília)	29.55%	**40.00%**	—
	Avg. RMSE / 100k	27.43	**26.80**	—
	Avg. SMAPE	**29.99%**	28.08%	—
	Avg. MDA	13.27%	**22.35%**	—
Non-Cleaned (All Cities)	RMSE (Brasília)	532.37	**528.25**	—
	SMAPE (Brasília)	59.43%	**57.72%**	—
	MDA (Brasília)	29.55%	**39.55%**	—
	Avg. RMSE / 100k	63.71	**62.78**	—
	Avg. SMAPE	28.91%	**19.41%**	—
	Avg. MDA	6.62%	**9.99%**	—

Table 6 Comprehensive evaluation of GCRN, LTGCN, and GTGCN across three data subsets: Top 40 Cities, Cleaned Cities, and Non-Cleaned (All Cities). Bolded values indicate best performance per row. Global attention was only feasible on the Top 40 due to memory constraints.

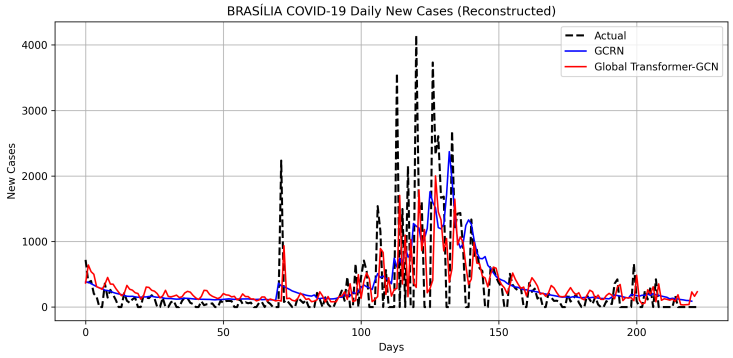

Fig. 4 Forecast comparison on BRASÍLIA: GCRN vs. GTGCN

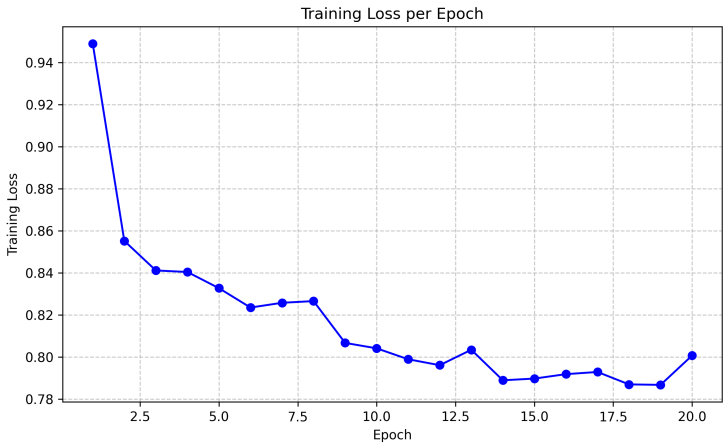

Fig. 5 Training loss graph for the GCRN on the Brazilian dataset, 20 Epochs

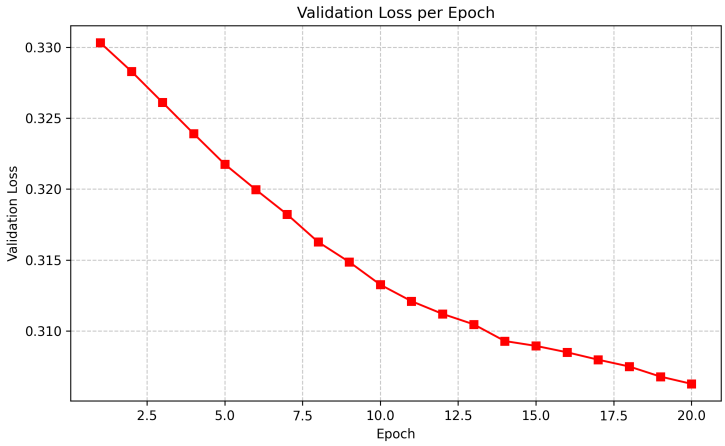

Fig. 6 Validation loss graph for the GCRN on the Brazilian dataset, 20 Epochs

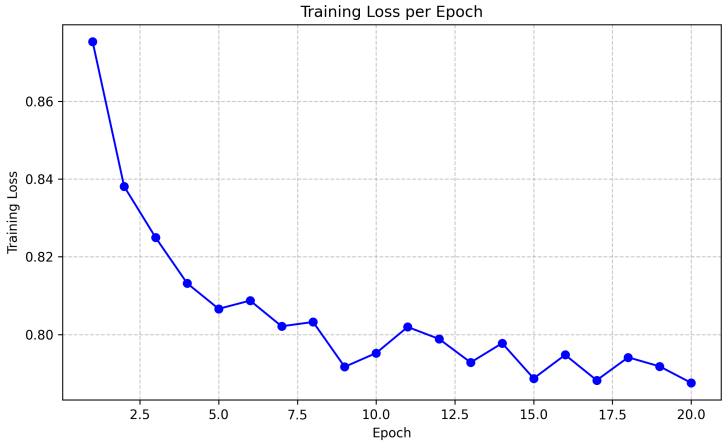

Fig. 7 Training loss graph for the LTGCN on the Brazilian dataset, 20 Epochs

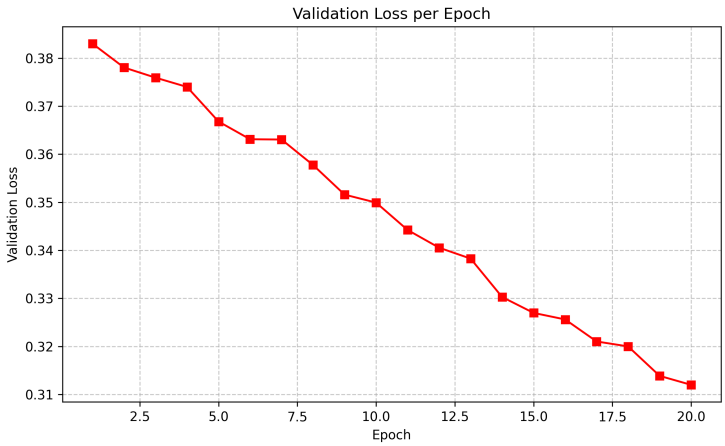

Fig. 8 Validation loss graph for the LTGCN on the Brazilian dataset, 20 Epochs

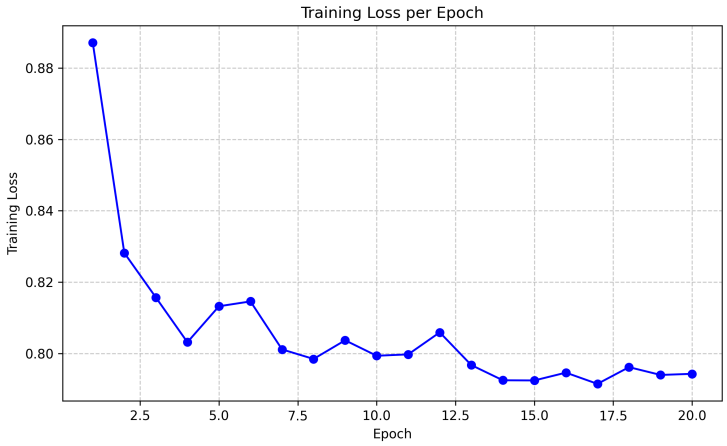

Fig. 9 Training loss graph for the GTGCN on the Brazilian dataset, 20 Epochs

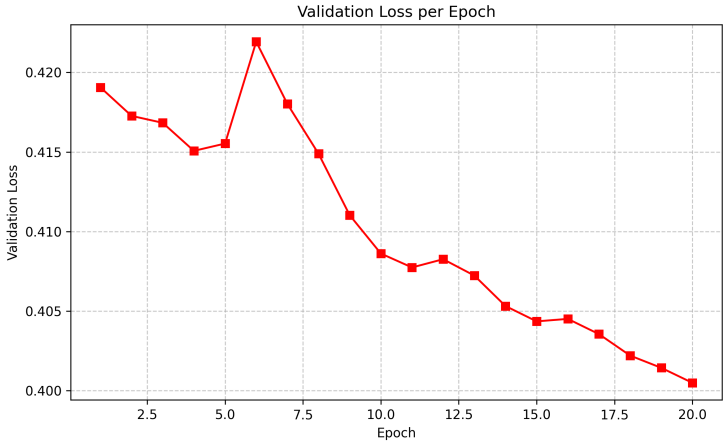

Fig. 10 Validation loss graph for the GTGCN on the Brazilian dataset, 20 Epochs

Overview of results (Brazil).

Figures 5 and 6 illustrate the training and validation dynamics of the *GCRN* over 20 epochs on the top 240 Cities in the Brazilian dataset. Figures 7 and 8 report analogous curves for the *LTGCN*, while Figures 9 and 10 summarize the performance of the *GTGCN*.

Table 7 Average training time per epoch on the Brazilian dataset.

Model	Avg. time / epoch (s)
GCRN (20 epochs)	40.70
LTGCN (20 epochs)	88.35
GTGCN (20 epochs)	235.07

Summary of Key Results

Several trends emerged from our experiments, which we explore in greater detail in the Discussion section. First, we observed a substantial performance gain across all models after cleaning the data, highlighting the importance of data quality. Second, the LTGCN consistently outperformed GCRN on directional accuracy metrics, showcasing the benefits of Transformer-based temporal modeling. Third, the GTGCN achieved the best overall results within the Top 40 city subset, demonstrating the utility of

cross-city attention despite its computational cost. Lastly, we found that RMSE can be biased toward high-population cities, making metrics like SMAPE and MDA more informative for regional comparisons, particularly during outbreak peaks.

4.5 Results on Spanish Dataset

Quantitative Results

Table 8 reports forecasting performance on the Spanish dataset across GCRN, LTGCN, and GTGCN. Metrics include RMSE, SMAPE, and MDA, evaluated for both Barcelona and the average across all 52 provinces including the standard deviation. Due to Spain's consistent case reporting and minimal anomalies, model performance is generally more stable than on the Brazilian dataset.

Subset	Metric	GCRN	LTGCN	GTGCN
Barcelona	RMSE	**3342.92**	3361.47	3620.19
	SMAPE	43.05%	**19.23%**	19.99%
	MDA	49.03%	62.99%	**64.29%**
All Provinces	Avg. RMSE / 100k	49.29 ± 6.22	**45.45 ± 5.62**	48.72 ± 6.94
	Avg. SMAPE	23.10 ± 4.07%	**17.86 ± 2.87%**	19.79 ± 3.86 %
	Avg. MDA	45.45 ± 6.54%	**55.13 ± 7.84%**	54.22 ± 6.67%

Table 8 Comparison of GCRN, LTGCN, and GTGCN on the Spanish dataset. Bold values indicate best performance per row.

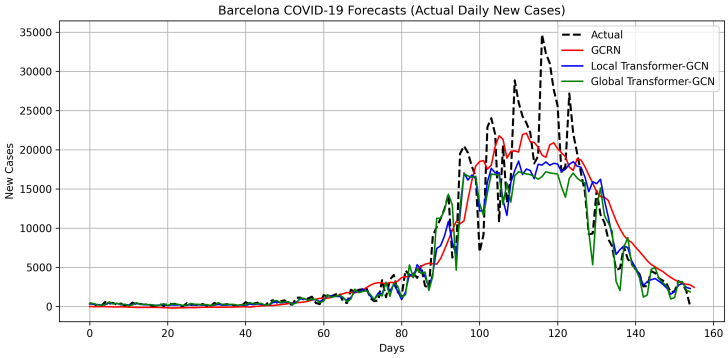

Fig. 11 Forecast comparison on BARCELONA across all models: GCRN, LTGCN, and GTGCN.

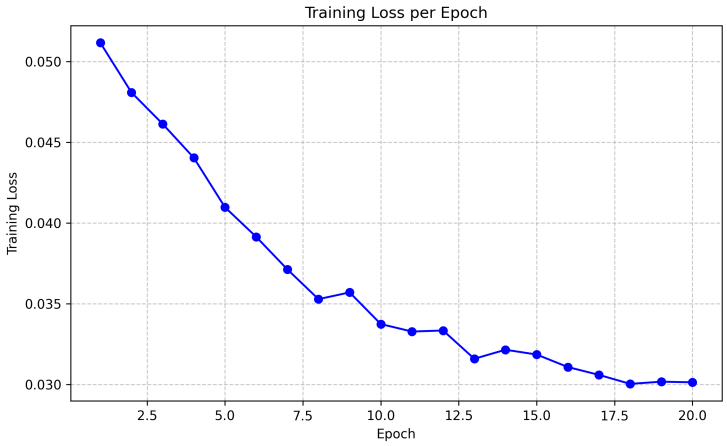

Fig. 12 Training loss graph for the GCRN on the spanish dataset, 20 Epochs

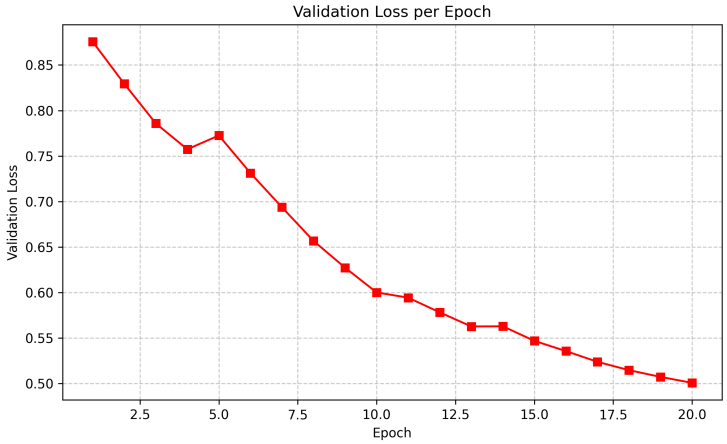

Fig. 13 Validation loss graph for the GCRN on the spanish dataset, 20 Epochs

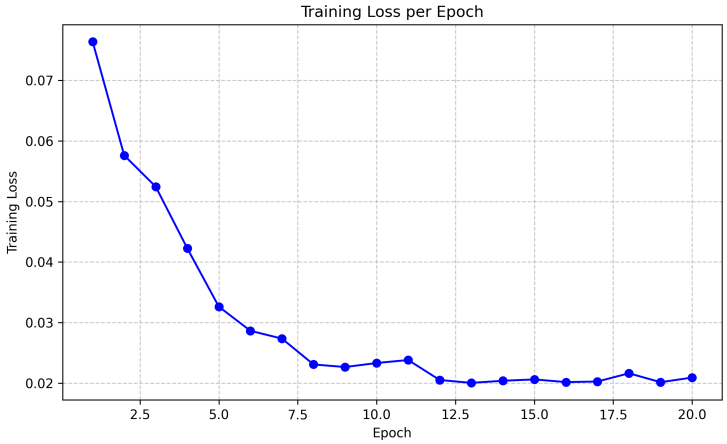

Fig. 14 Training loss graph for the LTGCN on the spanish dataset, 20 Epochs

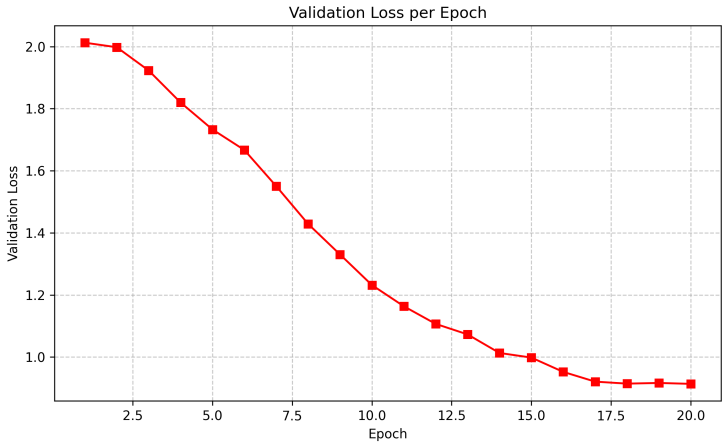

Fig. 15 Validation loss graph for the LTGCN on the spanish dataset, 20 Epochs

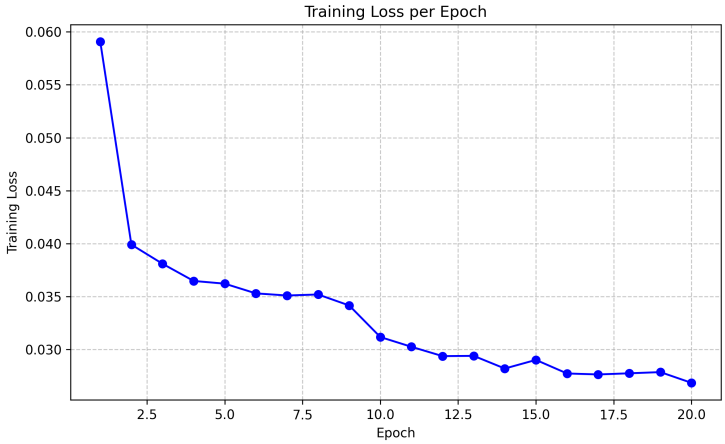

Fig. 16 Training loss graph for the GTGCN on the spanish dataset, 20 Epochs

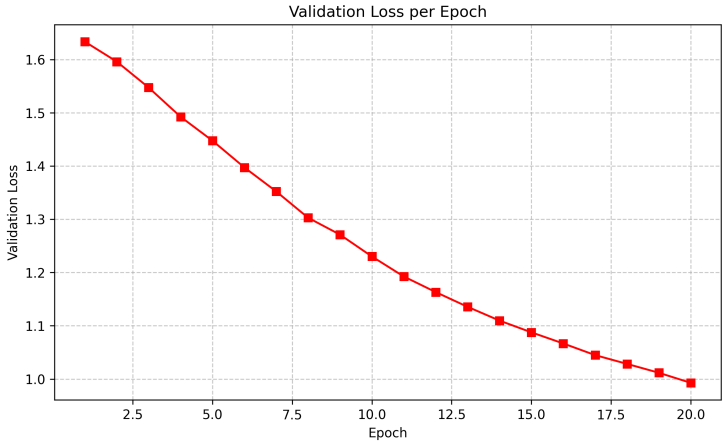

Fig. 17 Validation loss graph for the GTGCN on the spanish dataset, 20 Epochs

Overview of results.

Figure 11 contrasts the forecasts for **BARCELONA** across the three architectures: GCRN, LTGCN, and GTGCN, highlighting overall fit and residual structure on the

same target series. The subsequent panels report learning dynamics on the **Spanish** dataset: Figures 12 and 13 show the *GCRN* training and validation losses over 20 epochs; Figures 14 and 15 show the *LTGCN* counterparts over 20 epochs; and Figures 16 and 17 report the *GTGCN* curves over 20 epochs.

Table 9 Average training time per epoch on the Spanish dataset.

Model	Avg. time / epoch (s)
GCRN (20 epochs)	13.51
LTGCN (20 epochs)	33.02
GTGCN (20 epochs)	113.17

Summary of Spanish Results

Compared to Brazil, Spain's uniformly reported dataset required no data cleaning, enabling models to perform optimally out of the box. GCRN served as a stable baseline but was outperformed by both Transformer-based models in all evaluation metrics. LTGCN achieved the best overall results, benefiting from nodewise attention without inter-node communication. GTGCN performed comparably on directional accuracy (MDA) but was less effective on RMSE and SMAPE, raising questions about its added complexity in homogeneously reported datasets. A deeper discussion of these results follows in the next section.

5 Discussion

This section synthesizes the results from both Brazil and Spain to draw broader conclusions about the predictive performance of spatio-temporal models in epidemic forecasting. We analyze the influence of data quality, architectural decisions (recurrent vs. Transformer-based), and spatial modeling approaches (GCN vs. global attention) on key evaluation metrics. Insights from the ablation study are integrated to understand model behavior under different configurations and input regimes.

Brazil: Data Quality, Transformer Power, and Population Bias

Data Cleaning Was Essential.

In the Brazilian dataset, moving from the full set of 5300+ cities to a cleaned subset of 1305 then 40 high-quality cities led to significant gains in forecasting accuracy. GCRN's average RMSE dropped from 63.71 to 27.43, and its MDA more than doubled. This highlights the importance of preprocessing: noisy reporting, prolonged zero-case streaks, and low-variance time series can severely obscure learning signals.

Transformers Outperform Recurrent Models.

LTGCN consistently outperformed GCRN in both SMAPE and MDA, across the full and cleaned datasets. Even without global attention, its nodewise temporal attention enabled better tracking of directional changes and outbreak shifts. These results demonstrate the flexibility and accuracy of Transformer-based models in handling epidemic time series with inconsistent reporting.

Global Attention Benefits High-Density Cities.

In the Top 40 most populous cities, GTGCN outperformed all other models in RMSE and SMAPE. Its ability to capture inter-city temporal dependencies made it well suited to synchronized outbreaks. However, this came at a computational cost: its

quadratic attention complexity ($\mathcal{O}(N^2)$) limits its scalability to larger graphs unless approximation techniques are introduced.

Population Skews RMSE.

A key observation from Brazil is the RMSE's sensitivity to population size. Cities like Brasília, with very high case counts, dominated the RMSE metric. In contrast, SMAPE and MDA, being scale-independent and trend-aware, offered more balanced comparisons. As illustrated in Figure 4, trend alignment during outbreak peaks varied widely between models, further emphasizing the value of direction-sensitive metrics such as MDA.

Spain: Consistency Enables Clean Comparisons

Uniform Data Quality Enhances Fairness.

Unlike Brazil, the Spanish dataset exhibited uniformly reported high-quality data across all 52 provinces. No cities were filtered out and all models were evaluated on the full dataset. This consistency allowed for clearer comparisons between architectures without the confounding effects of data irregularities.

GCRN Serves as a Stable Baseline.

GCRN achieved an average RMSE of 49.29 per 100k, SMAPE of 23.10%, and MDA of 45.45%, serving as a solid baseline. It recorded the lowest RMSE in Barcelona (3342.92), but was outperformed overall by both Transformer-based variants, especially in proportional accuracy and trend tracking.

Local Attention Achieves the Best Overall Results.

LTGCN achieved the lowest average RMSE (45.45), lowest SMAPE (17.86%), and highest MDA (55.13%). Compared to GCRN, this reflects a 3.84 point RMSE reduction, a 22.7% improvement in SMAPE, and nearly 10 percentage point gain in MDA. Its nodewise attention mechanism was sufficient to model outbreak dynamics effectively in Spain's homogeneously reported setting.

Global Attention Offers Marginal Gains with Tradeoffs.

GTGCN achieved strong directional performance (MDA: 54.22%) and slight gains in peak awareness, especially in synchronized outbreaks like those in Barcelona. However, its average RMSE (48.72) and SMAPE (19.79%) lagged behind the local model. The added complexity of global attention appears less justified for compact, consistently reported graphs.

Ablation Study and Architectural Insights

We conducted a series of ablation experiments to evaluate the impact of model depth, sequence design, and spatial encoding in epidemic forecasting. Our findings underscore the critical role of architectural decisions in both baseline GCRN and Transformer-based models.

Enhancing GCRN Reactivity.

Increasing the GRU hidden dimension from 64 to 512 and shortening the input window from 14 to 7 days yielded substantial improvements in responsiveness. As shown in

Figure 18, Barcelona's RMSE dropped from 4387.22 to 3342.92. These modifications allowed the model to better capture recent dynamics and rapid outbreak surges.

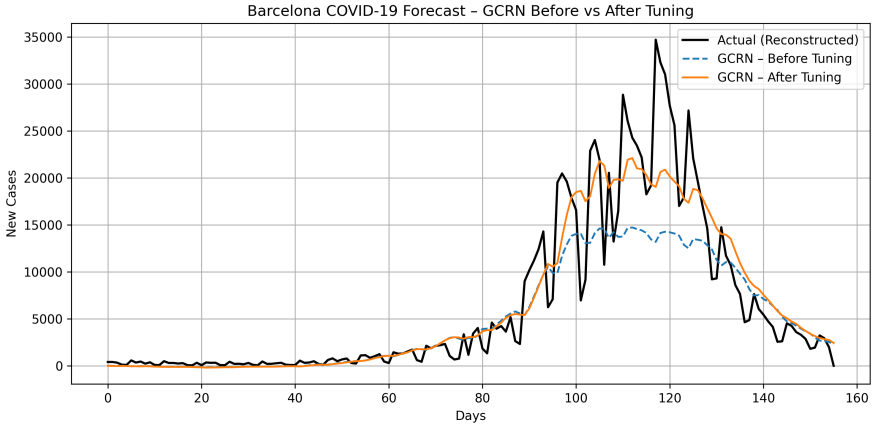

Fig. 18 Barcelona COVID-19 forecast: GCRN before vs. after tuning. The tuned model improves peak detection and magnitude alignment.

Role of Spatial Encoding.

To isolate the effect of spatial features, we removed the GCN encoder from the Transformer-GCN. While temporal trends were still captured, peak magnitudes were often misaligned. In Barcelona, the second peak was underestimated (Figure 19), and RMSE increased to 3476.19. On average, RMSE and SMAPE worsened slightly, while MDA unexpectedly increased, suggesting that GCN smoothing may trade off short-term directional precision for improved magnitude accuracy.

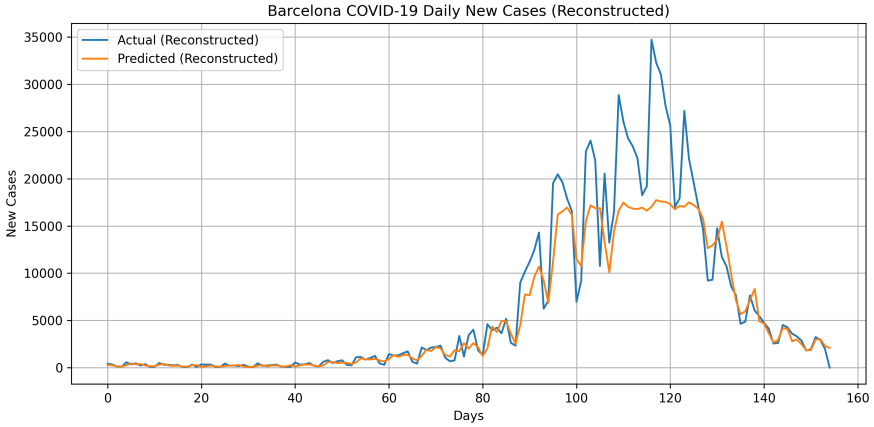

Fig. 19 Forecasts without GCN encoding led to peak underestimation and poorer scale calibration in Barcelona.

Summary of Insights.

These experiments offer key takeaways:

- **Shorter input sequences** might enhance reactivity in recurrent models.
- **GCN spatial context** improves magnitude estimation and inter-city calibration.
- **Hybrid architectures** combining GCN and Transformer modules consistently outperform their isolated counterparts across most metrics.

Limitations and Future Work

Limitations

Despite the promising performance of our Transformer-GCN models, several limitations must be acknowledged regarding data quality, modeling design, and evaluation.

Data Quality and Reporting Noise.

The Brazilian dataset contains irregular reporting patterns, such as prolonged zero-case streaks followed by large spikes and occasional negative values. These artifacts, likely due to administrative delays or retroactive corrections, introduce noise that may affect model learning. Although filtering and clipping were applied, residual distortions remain. To counterbalance this, we included a more stable dataset from Spain to evaluate robustness under cleaner reporting conditions.

Forecasting Setup Constraints.

All models were trained with fixed prediction horizon of one day and city-wise z-score normalization. While these choices simplify the learning task, they may limit the capture of long-range temporal dependencies and introduce distortions during highly non-stationary phases.

Evaluation Limitations.

Although RMSE was the primary evaluation metric, it is sensitive to population size and does not fully capture the alignment in outbreak timing. We supplemented this with SMAPE and MDA, yet qualitative aspects, such as trend sharpness and peak timing, remain partially unquantified.

Future Work

Addressing the above limitations suggests promising directions for future research:

- **Hybrid architectures:** Combining Transformers with sequential smoothing units such as GRUs [40] may better balance long-range context and local continuity.
- **Sparse attention for scalability:** Employing sparse or linear attention mechanisms [2, 53] can reduce computational cost and enable longer forecasting horizons.
- **Multi-modal temporal fusion:** Integrating auxiliary signals such as climate data, sentiment trends, or containment measures could improve robustness and interpretability.
- **Transfer learning:** Leveraging cross-region training and fine-tuning may enhance generalization to data-scarce or unseen locations.

Pursuing these avenues will strengthen the interpretability, accuracy, and scalability of spatio-temporal DL systems for epidemic forecasting and related real-world applications.

6 Conclusion

This work investigated the effectiveness of advanced spatio-temporal DL architectures for regional epidemic forecasting, focusing on hybrid models that combine GNNs with Transformer-based temporal encoders. Evaluated across COVID-19 datasets from

Brazil and Spain, these models were assessed for their ability to capture both spatial dependencies and complex temporal dynamics.

Our findings highlight LTGCN as a strong all-around performer, offering a favorable trade-off between accuracy, robustness, and scalability. GTGCN achieved the highest overall accuracy (in terms of RMSE and SMAPE but incurred substantial computational overhead due to its quadratic attention complexity. Notably, Transformer-based models consistently outperformed traditional recurrent baselines in MDA, confirming their ability to detect trend reversals, an essential property for early-warning systems and public health response.

Additionally, we demonstrate the importance of rigorous preprocessing strategies, including population normalization, data quality filtering, and spatial backbone construction. Integrating GCN-based encoders improved forecast calibration across regions, particularly in capturing the timing and magnitude of outbreak peaks.

Overall, our approach contributes to a flexible and scalable forecasting framework suitable for diverse public health applications, including regional alert systems, resource planning, and real-time epidemic monitoring. By fusing temporal attention, spatial graph reasoning, and robust data handling, our models form a strong foundation for next-generation epidemic forecasting systems.

Declarations

- The authors declare no conflict of interest.
- Source code is available at https://github.com/mobyous/LTGCN-GTGCN

References

[1] Can Rong, Jingtao Ding, and Yong Li. "An Interdisciplinary Survey on Origin-destination Flows Modeling: Theory and Techniques". In: *ACM Computing Surveys* 57.1 (Oct. 2024), pp. 1–49. ISSN: 1557-7341. DOI: 10.1145/3682058. URL: http://dx.doi.org/10.1145/3682058.

[2] Haoyi Zhou et al. "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting". In: *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*. 2021, pp. 11106–11115. DOI: 10.1609/AAAI.V35I12.17325.

[3] Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. "TranAD: deep transformer networks for anomaly detection in multivariate time series data". In: *Proc. VLDB Endow.* 15.6 (Feb. 2022), pp. 1201–1214. ISSN: 2150-8097. DOI: 10.14778/3514061.3514067. URL: https://doi.org/10.14778/3514061.3514067.

[4] Cornelius Fritz, Emilio Dorigatti, and David Rügamer. "Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany". In: *Sci. Rep.* 12.1 (Mar. 2022), p. 3930.

[5] Yulan Li, Yang Wang, and Kun Ma. "Integrating Transformer and GCN for COVID-19 Forecasting". In: *Sustainability* 14.16 (2022). ISSN: 2071-1050. DOI: 10.3390/su141610393. URL: https://www.mdpi.com/2071-1050/14/16/10393.

[6] Ming Jin et al. " A Survey on Graph Neural Networks for Time Series: Forecasting, Classification, Imputation, and Anomaly Detection ". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 46.12 (Dec. 2024), pp. 10466–10485. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2024.3443141. URL: https://doi.ieeecomputersociety.org/10.1109/TPAMI.2024.3443141.

[7] Shiyu Yang et al. "SSGCRTN: a space-specific graph convolutional recurrent transformer network for traffic prediction". In: *Applied Intelligence* 54.22 (2024), pp. 11978–11994. ISSN: 1573-7497. DOI: 10.1007/s10489-024-05815-1. URL: https://doi.org/10.1007/s10489-024-05815-1.

[8] Shiyu Yang et al. "MSTDFGRN: A Multi-view Spatio-Temporal Dynamic Fusion Graph Recurrent Network for traffic flow prediction". In: *Computers and Electrical Engineering* 123 (2025), p. 110046. ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2024.110046. URL: https://www.sciencedirect.com/science/article/pii/S0045790624009716.

[9] Liu Aoyu and Yaying Zhang. "Spatial–Temporal Dynamic Graph Convolutional Network With Interactive Learning for Traffic Forecasting". In: *IEEE Transactions on Intelligent Transportation Systems* PP (July 2024), pp. 1–16. DOI: 10.1109/TITS.2024.3362145.

[10] Shiyu Yang and Qunyong Wu. "SDSINet: A spatiotemporal dual-scale interaction network for traffic prediction". In: *Appl. Soft Comput.* 173 (2025), p. 112892. URL: https://api.semanticscholar.org/CorpusID:276572527.

[11] Feiyan Sun et al. "TVGCN: Time-varying graph convolutional networks for multivariate and multifeature spatiotemporal series prediction". In: *Science Progress* 107.3 (2024), p. 00368504241283315. DOI: 10.1177/00368504241283315. URL: https://doi.org/10.1177/00368504241283315.

[12] Leonardo López and Xavier Rodó. "A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: Simulating control scenarios and multi-scale epidemics". In: *Results in Physics* 21 (2021), p. 103746. ISSN: 2211-3797. DOI: https://doi.org/10.1016/j.rinp.2020.103746. URL: https://www.sciencedirect.com/science/article/pii/S2211379720321604.

[13] Serina Chang et al. "Mobility network models of COVID-19 explain inequities and inform reopening". In: *Nature* 589.7840 (Nov. 2020), pp. 82–87. DOI: 10.1038/s41586-020-2923-3.

[14] Giulia Giordano et al. "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy". en. In: *Nat. Med.* 26.6 (June 2020), pp. 855–860.

[15] Duygu Balcan et al. "Multiscale mobility networks and the spatial spreading of infectious diseases". In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21484–21489. DOI: 10.1073/pnas.0906910106. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0906910106.

[16] Matteo Chinazzi et al. “The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak”. In: *Science* 368.6489 (2020), pp. 395–400. DOI: 10.1126/science.aba9757. URL: https://www.science.org/doi/abs/10.1126/science.aba9757.

[17] Qi Deng. “Dynamics and development of the COVID-19 epidemic in the United States: A compartmental model enhanced with deep learning techniques”. In: *J. Med. Internet Res.* 22.8 (Aug. 2020), e21173.

[18] Junaid Farooq and Mohammad Abid Bazaz. “A deep learning algorithm for modeling and forecasting of COVID-19 in five worst affected states of India”. In: *Alexandria Engineering Journal* 60.1 (2021), pp. 587–596. ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2020.09.037. URL: https://www.sciencedirect.com/science/article/pii/S1110016820304907.

[19] Xiao Ning et al. “Epi-DNNs: Epidemiological priors informed deep neural networks for modeling COVID-19 dynamics”. en. In: *Comput. Biol. Med.* 158.106693 (May 2023), p. 106693.

[20] Philip Nadler, Rossella Arcucci, and Yike Guo. “A Neural SIR Model for Global Forecasting”. In: *Proceedings of the Machine Learning for Health NeurIPS Workshop*. Vol. 136. Proceedings of Machine Learning Research. PMLR, Nov. 2020, pp. 254–266. URL: https://proceedings.mlr.press/v136/nadler20a.html.

[21] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[22] Sourabh Shastri et al. “Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study”. In: *Chaos Solitons Fractals* 140.110227 (Nov. 2020), p. 110227.

[23] Khondoker Nazmoon Nabi et al. “Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks”. In: *Results in Physics* 24 (2021), p. 104137. ISSN: 2211-3797. DOI: https://doi.org/10.1016/j.rinp.2021.104137. URL: https://www.sciencedirect.com/science/article/pii/S2211379721002904.

[24] Mohamed Marzouk et al. “Deep learning model for forecasting COVID-19 outbreak in Egypt”. In: *Process Safety and Environmental Protection* 153 (2021), pp. 363–375. ISSN: 0957-5820. DOI: https://doi.org/10.1016/j.psep.2021.07.034. URL: https://www.sciencedirect.com/science/article/pii/S0957582021004055.

[25] Xiaofeng Zhu et al. “Modeling epidemic dynamics using Graph Attention based Spatial Temporal networks”. In: *PLOS ONE* 19.7 (July 2024), pp. 1–22. DOI: 10.1371/journal.pone.0307159. URL: https://doi.org/10.1371/journal.pone.0307159.

[26] Junkai Mao, Yuexing Han, and Bing Wang. “MPSTAN: Metapopulation-Based Spatio–Temporal Attention Network for Epidemic Forecasting”. In: *Entropy* 26.4 (2024). ISSN: 1099-4300. DOI: 10.3390/e26040278. URL: https://www.mdpi.com/1099-4300/26/4/278.

[27] George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. “Transfer Graph Neural Networks for Pandemic Forecasting”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.6 (May 2021), pp. 4838–4845. DOI: 10.1609/aaai.v35i6.16616. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16616.

[28] F. Duarte, G. Moreira, and V. Freitas. “Graph neural networks for COVID-19 spread prediction using mobility data”. In: *IEEE Access* 10 (2022), pp. 85494–85506.

[29] Fernando Henrique Duarte et al. *Leveraging graph neural networks and mobility data for COVID-19 forecasting*. Jan. 2025. DOI: 10.48550/arXiv.2501.11711.

[30] Instituto Brasileiro de Geografia e Estatística (IBGE). *Survey on the Road and Waterway Intercity Mobility Network*. Available from https://www.ibge.gov.br. 2017.

[31] Movilidad y Agenda Urbana Ministerio de Transportes. *Movilidad en las comunidades autónomas: Datos de movilidad*. Available from https://www.mitma.gob.es/. 2023.

[32] Datadista. *COVID-19 datasets - Spain*. Accessed: 2025-05-03. 2020. URL: https://github.com/datadista/datasets/blob/master/COVID%2019/readme.md.

[33] M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. "Extracting the multiscale backbone of complex weighted networks". In: *Proceedings of the National Academy of Sciences* 106.16 (Apr. 2009), pp. 6483–6488. ISSN: 1091-6490. DOI: 10.1073/pnas.0808904106. URL: http://dx.doi.org/10.1073/pnas.0808904106.

[34] Matthias Fey and Jan Eric Lenssen. *Fast Graph Representation Learning with PyTorch Geometric*. 2019. arXiv: 1903.02428 [cs.LG]. URL: https://arxiv.org/abs/1903.02428.

[35] Youngjoo Seo et al. "Structured Sequence Modeling with Graph Convolutional Recurrent Networks". In: *Neural Information Processing*. Ed. by Long Cheng, Andrew Chi Sing Leung, and Seiichi Ozawa. Cham: Springer International Publishing, 2018, pp. 362–373. ISBN: 978-3-030-04167-0.

[36] Yaguang Li et al. "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=SJiHXGWAZ.

[37] Zonghan Wu et al. "Graph WaveNet for Deep Spatial-Temporal Graph Modeling". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 1907–1913. DOI: 10.24963/ijcai.2019/264. URL: https://doi.org/10.24963/ijcai.2019/264.

[38] Yanbei Liu et al. "Structural Attention Graph Neural Network for Diagnosis and Prediction of COVID-19 Severity". In: *IEEE Transactions on Medical Imaging* PP (Dec. 2022), pp. 1–1. DOI: 10.1109/TMI.2022.3226575.

[39] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[40] Shiyang Li et al. "Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/6775a0635c302542da2c32aa19d86be0-Paper.pdf.

[41] Haixu Wu et al. "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 22419–22430. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/bcc0d400288793e8bdcd7c19a8ac0c2b-Paper.pdf.

[42] George Zerveas et al. *A Transformer-based Framework for Multivariate Time Series Representation Learning*. 2021. URL: https://openreview.net/forum?id=lE1AB4stmX.

[43] Mingxing Xu et al. *Spatial-Temporal Transformer Networks for Traffic Flow Forecasting*. 2021. arXiv: 2001.02908 [eess.SP]. URL: https://arxiv.org/abs/2001.02908.

[44] Zhengyu Li, Hongjie Zhang, and Wei Zheng. "STformer: Spatio-Temporal Transformer for Multivariate Time Series Anomaly Detection". In: *Artificial Neural Networks and Machine Learning – ICANN 2024*. Ed. by Michael Wand et al. Cham: Springer Nature Switzerland, 2024, pp. 297–311. ISBN: 978-3-031-72347-6.

[45] Junyi Gao et al. "STAN: spatio-temporal attention network for pandemic prediction using real-world evidence". In: *Journal of the American Medical Informatics Association* 28.4 (Jan. 2021), pp. 733–743. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa322. eprint: https://academic.oup.com/jamia/article-pdf/28/4/733/36642145/ocaa322.pdf. URL: https://doi.org/10.1093/jamia/ocaa322.

[46] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[47] Chengxuan Ying et al. "Do Transformers Really Perform Badly for Graph Representation?" In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 28877–

28888. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/f1c1592588411002af340cbaedd6fc33-Paper.pdf.

[48] Cunjun Yu et al. *Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction*. 2020. arXiv: 2005.08514 [cs.CV]. URL: https://arxiv.org/abs/2005.08514.

[49] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[50] Firuz Kamalov et al. "Deep learning for Covid-19 forecasting: State-of-the-art review." In: *Neurocomputing* 511 (2022), pp. 142–154. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2022.09.005. URL: https://www.sciencedirect.com/science/article/pii/S0925231222010918.

[51] Nooshin Ayoobi et al. "Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods". In: *Results in Physics* 27 (2021), p. 104495. ISSN: 2211-3797. DOI: https://doi.org/10.1016/j.rinp.2021.104495. URL: https://www.sciencedirect.com/science/article/pii/S2211379721006069.

[52] Silvino Pedro Cumbane and Győző Gidófalvi. "Deep learning-based approach for COVID-19 spread prediction". In: *International Journal of Data Science and Analytics* (2024). DOI: 10.1007/s41060-024-00558-1. URL: https://doi.org/10.1007/s41060-024-00558-1.

[53] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. 2020. arXiv: 2004.05150 [cs.CL]. URL: https://arxiv.org/abs/2004.05150.